Amendments to the Specification

Paragraph from page 4, lines 21-23

Viseme- the minimum distinctive visual manifestation of an acoustic identification (e.g., of an articulatory type) [[.]] Representation in a video or motion picture.

Paragraph from page 5, line 1- page 6, line 5

— The present invention takes advantage of the advancements achieved in the field of visual information, or visemes, in speech recognition, which are the subject of co-pending U.S. Patent application Serial No. 09/452,919 filed December 2, 1999 (YO999-428) entitled "Late Integration in Audio-Visual Continuous Speech Recognition" by Verma, et al; patent application Serial No: 09/369,707 (YO999-317) entitled "Methods and Apparatus for Audio-Visual Speech Detection and Recognition" by S. Basu, et al; and Serial No: 09/369,706 (YO999-318) entitled "Methods and Apparatus for Audio-Visual Speaker Recognition and Utterance Verification" by S. Basu, et al now U. S. Patent No. 6,219,640 which issued on 17 April 2001. As detailed therein, visual information, such as the mouth parameters of height, width, and area, along with derivative image information are used to continuously recognize speech, particularly in a non-controlled environment which may have multiple extraneous noise sources. Further to the enhancement of speech recognition using facial analysis (see: the 6,219,640 patent ~~09/369,707 application~~) and the speaker recognition using audio and visual recognition techniques (the ~~09/369,706~~ 6,219,640 patent ~~application~~), the Verma patent application focusses on the fusion (or alignment) of data output from a visual recognizer and audio recognizer to improve speech recognition accuracy and to

provide automatic speech detection. More particularly, the Verma patent application processes a video signal to identify a class of the most likely visemes found in the signal. Thereafter, the most likely phones and/or phonemes associated with the identified visemes, or with the audio signal, are considered for audio recognition purposes. Therefore, the system and method of the Verma patent applications use both audio and video processing to discern phones produced by the subject, and the phones are, in turn, linked together to discern words.

Paragraph from page 14, line 4-page 15, line 2

It is noteworthy that the synchronization algorithm can be applied, as desired, to prerecorded audiovisual materials; or it can be applied on-the-fly and continuously to, for example, "live" audiovisual materials. Also, although this invention has been described using English-language examples, there is nothing that restricts it to English and it can be implemented for any language. Finally, it should be understood that, while the highest visual recognition accuracy has been realized using facial features linked to speech, it is possible to recognize non-speech acoustic signatures, to link those non-speech acoustic signatures to non-speech visual "cues" (for example, hand-clapping), to time-stamp the audio and visual output streams, and to synchronize the audio and video based on the identified cues in the time stamped output streams. Under such a visual recognition scenario, the process flow of Fig. 2 would be generalized to the steps of image extraction, feature detection, feature parameter analysis, and correlation of acoustic signatures stored in a database to the feature parameters. For a detailed discussion of the training and use of speech recognition means for identifying audio sources by acoustic signatures, please see co-pending patent application Serial No: 09/602,452, (YOR9-2000-0130) entitled "System and Method for Control of Lights, Signals, Alarms Using Sound Detection" by W. Ablondi, et al, the teachings of which are herein incorporated by reference.